Chapter 5

Genotyping of complex lesions

5.1 Introduction

Most PCR-based genotyping of Cas9-induced lesions has so far focused on the region immediately adjacent to the cut site (<1000 bp) in bulk cell populations. This biases the assessment by excluding lesions that destroy the primer binding sites (large deletions), disconnect them (translocation and large inversions) or prevent amplification by increasing the distance between them (large insertions). Cas9-induced lesions that are non-contiguous with the cut site and outside of the amplified region are also missed. Failure to recover such complex alleles is not apparent, when genotyping in bulk cell populations. While some specialized, PCR-based methods for detection of such lesions in bulk populations exist, they are not broadly used. PCR employed by these methods also has to be anchored in one flank of of the break, which biases the output (Cain-Hom et al., 2017; de Vree et al., 2014; Giannoukos et al., 2018; Zheng et al., 2014).

Here, I addressed some of these issues by combining long-range PCR with Sanger and PacBio sequencing to detect and describe complex Cas9-induced lesions in an unbiased way.

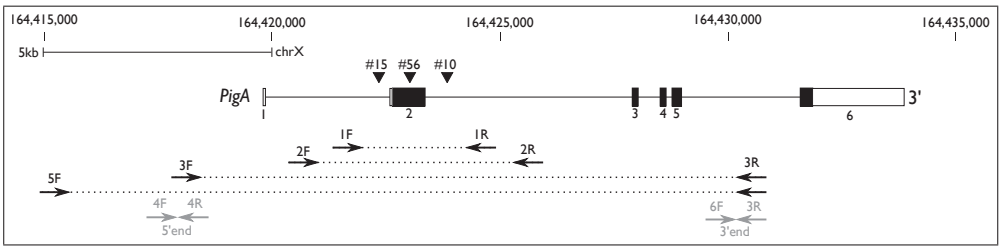
5.2 Results

5.2.1 Deletions underlying loss of gene expression caused by intronic gRNAs

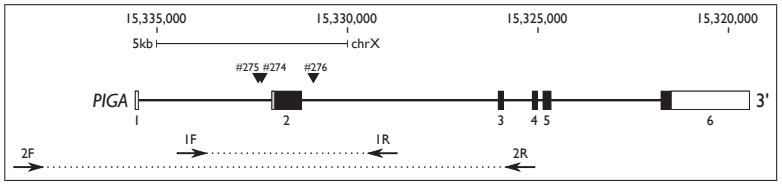
In chapter 4, I showed that individual, intronic gRNAs can cause loss of expression of the chromosome X linked PigA in about 12% of transfected male mouse ES cells. To understand what genetic changes underlie this phenotype, I amplified a 5.7 kb region around exon 2 from pools of cells mutagenized with three selected gRNAs introduced by PiggyBac transposition and sequenced the PCR products using the PacBio platform. I observed a depletion in read coverage on a kilobase-scale around the cut sites, consistent with the presence of large deletions (Fig. 5.2). Cells mutagenized with intronic guides and sorted for loss of PigA generally exhibited loss of the adjacent exon 2 (Fig. 5.1a). If intronic regulatory sequences were present around the exon, the DNA of cells sorted for retention of PigA expression would be wild type or contain only small indels around the cut site. However, the most frequent lesions in these cells were kilobase-scale deletions extending away from the exon. I conclude that, in most cases, loss of PigA expression was likely caused by loss of the exon, rather than damage to intronic regulatory elements.

PacBio sequencing of pooled edited DNA is biased towards detection of large deletions. PCR is more likely to amplify shorter amplicons, favoring deletions. Capture of short fragments is also more efficient during PacBio sequencing. Finally, individual, shorter DNA molecules are usually read more times during sequencing than longer ones. As a consequence, they have higher quality scores and are more likely to pass quality filters. Another disadvantage of the PacBio approach is the need to choose the amplicon size beforehand, which means some alleles with larger lesions may be missed. Finally, translocations cannot be amplified by a pair of fixed primers at all.

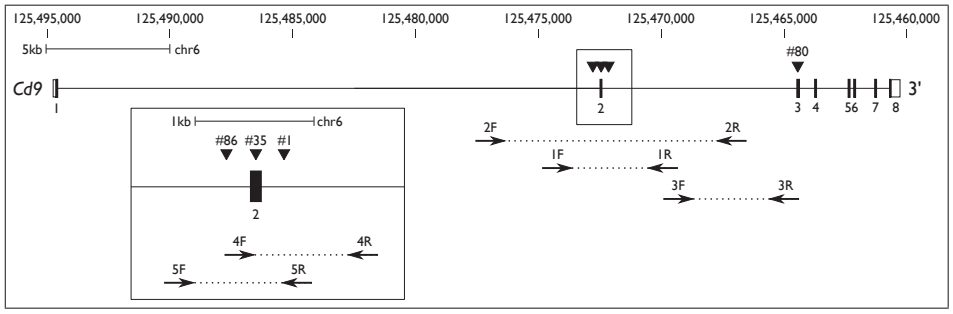
Therefore, to fully characterize the variety of mutagenized alleles, I isolated single cell clones. The loci around the gRNA target site were ampli-



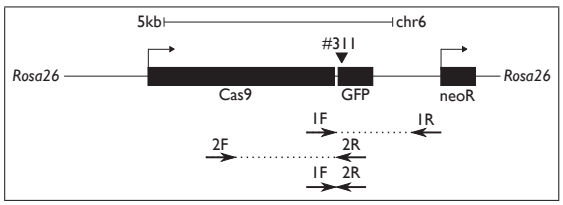
(a) *PigA* locus.



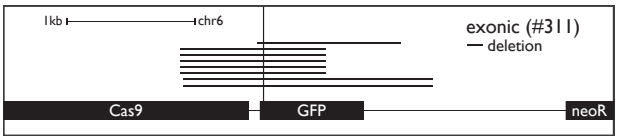
(b) *PIGA* locus.



(c) *Cd9* locus.



(d) *Cas9-GFP* locus, bone marrow progenitors experiment.



(e) Alleles recovered by Sanger sequencing from mouse bone marrow progenitor cells mutagenized at the *Cas9-GFP* locus. The position of the gRNA #311 is shown as a vertical line intersecting with the gene structure. Horizontal line indicates deletion.

Figure 5.1: Positions of primer pairs and gRNAs used for genotyping experiments (Table 5.1 and 5.3). Genomic position is given with respect to the GRCm38 or GRCh38 reference genome. In grey, primers used for diagnostic PCRs on alleles that could not be recovered.

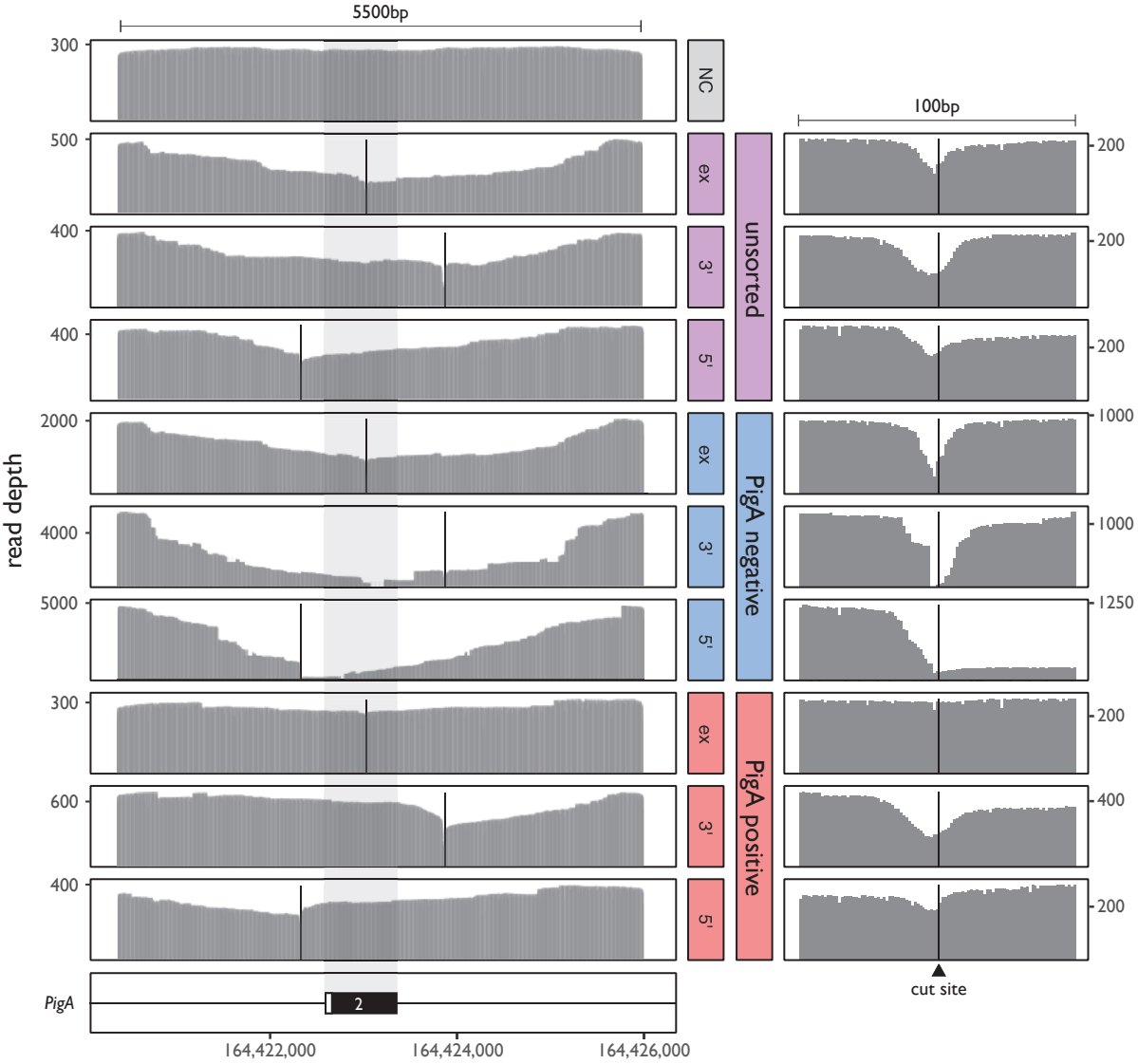
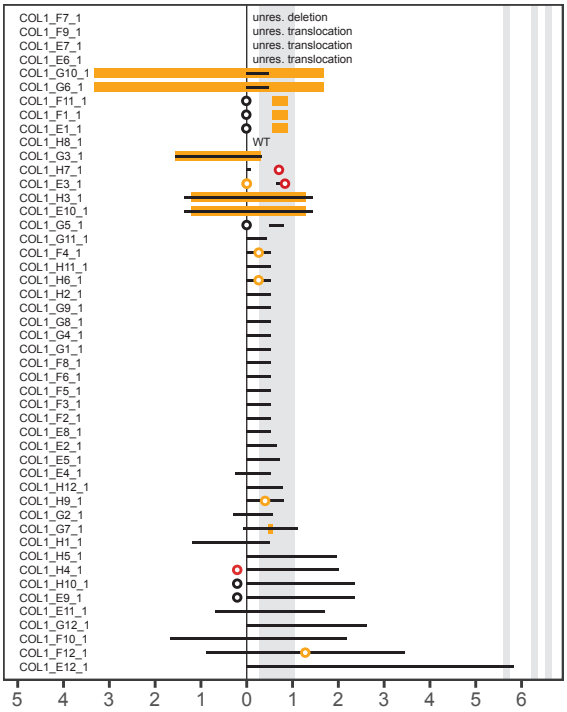
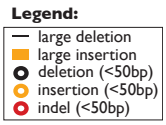
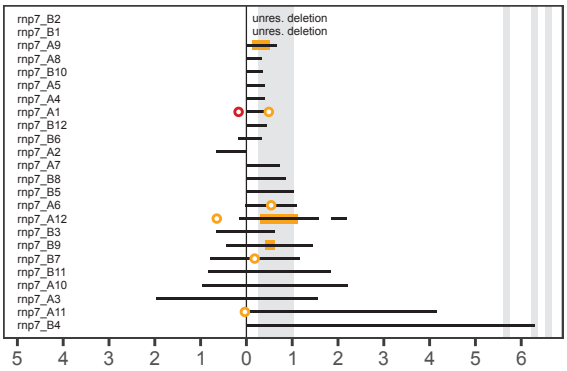


Figure 5.2: Analysis of the *PigA* locus mutagenized with selected gRNAs. Coverage of PacBio reads at the *PigA* locus. The locus was PCR-amplified from a pool of cells sorted for PigA expression (or from the unsorted population), and the resulting products were sequenced using the PacBio platform. The right panel depicts a 100 bp region centered at the cut site. NC: negative-control gRNA, ex: exonic gRNA (#56), 5' : 5' intronic gRNA (#15), 3' : 3' intronic gRNA (#10). The cut site of the gRNA is indicated with a vertical black bar. Genomic position is given with respect to the GRCm38 reference genome. N = 1.

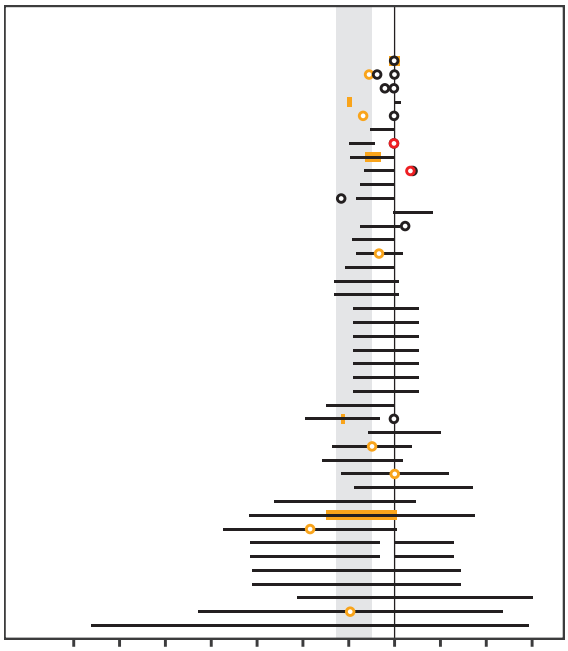


(a) 5' intronic guide #15, PiggyBac method.

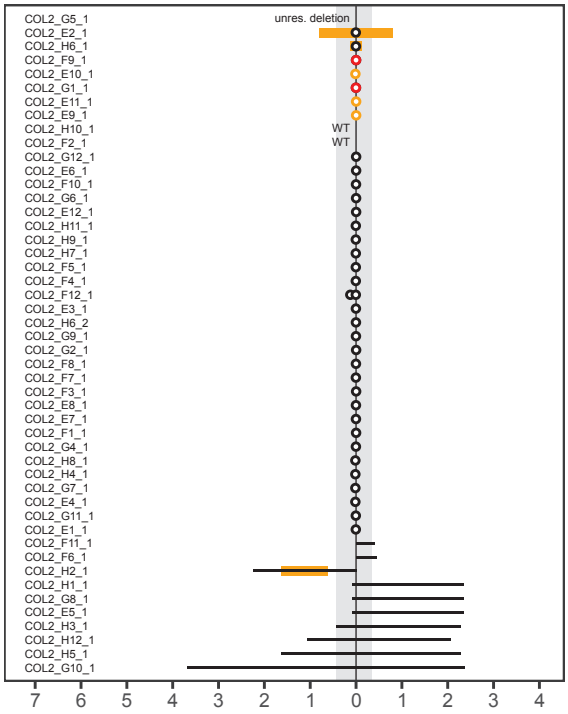




(b) 5' intronic guide #15, RNP method.

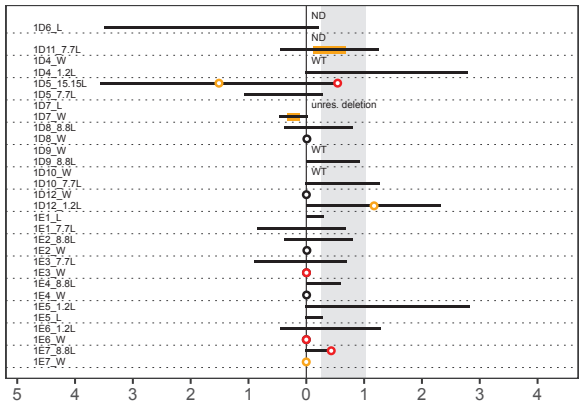


(c) 3' intronic guide #10, PiggyBac method.

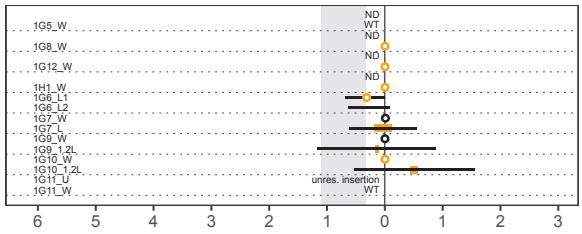


(d) Exonic guide #56, PiggyBac method.

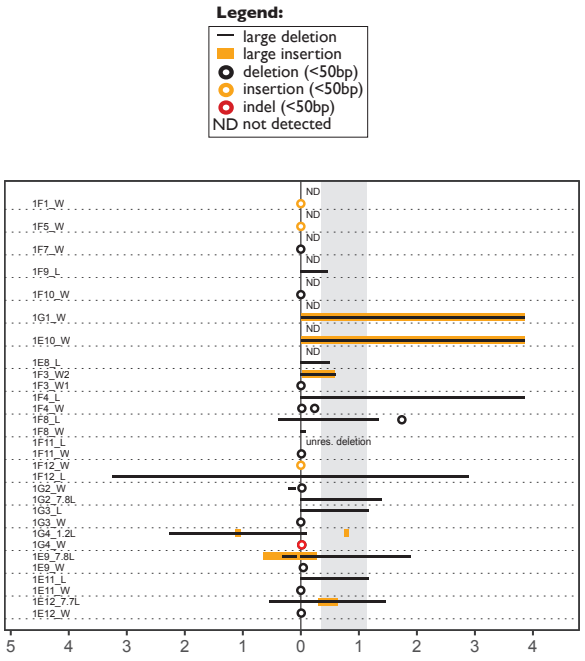
Figure 5.3: Alleles recovered by Sanger sequencing from Cas9-edited, PigA-deficient mouse ES cell clones. The position of the gRNA is shown as a vertical line. Pure insertions and deletions of <50 bp are indicated with orange and black circles, respectively. Combined insertion/deletion events of <50 bp and SNPs ("indel (<50 bp)" in the legend) are indicated with a red circle. Black lines represent deletions >50 bp. Orange bars indicate size of the >50 bp insertions (but not their map position). They are centered on the insertion locus or on the associated deletion. Gray shades represent exons 2 (large one), 3, 4 and 5. X-axis represents distance from the gRNA position in kilobases. Alleles are sorted by total length. Their names are indicated on the left.



(a) 5’ intronic guide #274.



(b) 3’ intronic guide #276.



(c) 5’ intronic guide #275.

Figure 5.4: Alleles recovered by Sanger sequencing from PIGA-deficient, human female RPE1 cell clones. Cas9-expressing cells were transiently transfected with gRNA-expressing plasmids. Grey shade represents exon 2. Names of individual alleles are indicated in the column on the left. Dotted horizontal line separates clones. Other display conventions as in Fig. 5.3.

fied using PCR primer pairs positioned progressively further apart (up to 12-16 kb), until amplicons were generated (see Fig. 5.1 for primers and gRNAs positions). Long elongation times were used to identify large insertions. Since mouse ES cells were grown on feeder cells (which help maintain their pluripotency), primer pairs which specifically exclude feeder cell DNA were used to avoid spurious wild-type alleles at the *PigA* locus. This could not be achieved at the *Cd9* locus due to low divergence between BL6 and feeder genomes. Amplicons were Sanger (all loci) or PacBio (only *Cd9* locus) sequenced. Since no wild-type CAST alleles were detected in any of the clones edited at the *Cd9* locus, I assumed all wild-type BL6 alleles in these clones were feeder derived.

Consistent with the results from PacBio sequencing, large deletions of >50 bp were detected in almost 85% (79/93) of PigA-deficient single cell clones generated by single, intronic gRNAs #10 and #15 (Fig. 5.3a and Fig. 5.3c). Most of them overlapped both the cut site and the nearest exon. The deletions varied in size, the largest spanning 9.5 kb. Identical results were obtained using electroporation of intronic gRNA #15 as RNP (Fig. 5.3b), as expected due to consistent rates of PigA-deficient cells between PiggyBac and RNP methods (chapter 4).

To assess the frequency of large deletions without strong selection for that outcome, I used the exonic guide #56 causing 97% PigA loss. Although two-thirds of alleles (32/48) from PigA-deficient cells had indels <50 bp, as expected, 20% (10/48) had deletions >50 bp, extending up to 6 kb. Some of the deletions exhibited clear directionality. Assuming this is also the case for deletions

induced by intronic guides, this explains why the observed rate of PigA loss with those guides is only ~12% (chapter 4), not 20%. It is also consistent with the depletion of read coverage distal to the exon in PacBio analysis (Fig. 5.2).

To replicate these results in another mouse ES cell line, I genotyped AB2.2 ES cell clones mutagenized at a hemizygous *GFP* or *mCherry* targeted transgene using the PiggyBac method. The lower frequency of deletions in these cells (7-12% vs the expected 20%, Table 5.1, “cherry/gfp” experiment)) is likely due to relatively shorter range of the PCR (<3 kb vs 16 kb). Consistently, no amplicon at all could be obtained in 15-26% of clones.

In chapter 4, I have shown that intronic gRNAs cause similar levels of PigA loss in both mouse ES cells and in human female differentiated RPE1 cell line. Consequently, I expected the rate of deletions in these cells to also be similar. I mutagenized RPE1 cells with intronic guides at the *PIGA* gene, isolated PIGA-deficient single cell clones and resolved their alleles using long-range PCR (up to 12 kb) and Sanger sequencing. Only deletions on the active chromosome X would be selected for in these cells, so I expected the rate of deletions on this chromosome to approximate 85% (as in PigA-deficient male ES cells mutagenized with intronic guides). Conversely, selection should not affect the inactive chromosome, resulting in the unselected rate of 20% (as in ES cells mutagenized with the exonic guide). The observed frequency of deletions in RPE1 cells is 47% (40/85), very close to the expected 51% (85%*0.5 + 20%*0.5; Fig. 5.4a, 5.4c and 5.4b). The largest deletion spanned 6 kb. All but four deletions overlapped both the cut site and the nearest exon.

Frequent deletions were also observed in mouse ES cells edited with intronic gRNA #1 at the bi-allelic *Cd9* locus. The rate of deletions was highest in the $Cd9^{low}$ (42/43) and lowest in "true wild-type" clones (17/55). Intermediate levels of deletions were observed in bimodal and "low turned wild-type" clones expressing intermediate levels of Cd9 (as defined in chapter 4, Fig. 5.5). See subsection 5.2.5 for a more in-depth discussion taking into account the allelic composition of individual clones.

To show that large deletions at the *Cd9* locus occur regardless of the choice of gRNA, I mutagenized the biallelic *Cd9* locus using two single intronic guides (#1 and #86; two replicates) and two single exonic guides (#35 and #80; one replicate), sorted for cells expressing different Cd9 levels, reassessed the expression of isolated clones and genotyped the clones using long-range PCR (Table 5.1, "cbbcs1" and "cbbcs3" experiments, Fig. 5.1c). In all examined groups a substantial fraction of clones had at least one deletion (18-88%). In particular, cells mutagenized with intronic guides and sorted for loss of gene expression collectively exhibit higher rates of deletion clones (50-88%) than cells sorted for retention of gene expression (33-46%). "Low turned wild-type" and bimodal clones exhibited an intermediate frequency of deletions (43-71%). In clones mutagenized with exonic gRNAs a large deletion is not necessary to ablate gene expression, as a small indel would have the same effect. Consequently, there was not clear correlation between clone expression level and fraction of deletion clones (range: 17-42%).

5.2.2 Deletions in primary bone marrow cells

Mouse ES cells can maintain pluripotency in culture for many passages. However, culture conditions could temporarily influence the DDR in these cells. Therefore, I replicated my results in primary cells. I chose to work with progenitor cells from the bone marrow of mice expressing Cas9-GFP from a transgene at the homozygous *Rosa26* locus. Lineage-negative cells enriched by removal of differentiated cells on magnetic columns were electroporated with a crRNA:trRNA complex against the *GFP* locus. GFP-negative single cell clones were isolated and genotyped around the cut site with three different primer pairs spanning in total 5 kb (Fig. 5.1d). At least one large deletion product between 100 bp and ~3 kb in size was detected in 36% of clones (35/96; Table 5.1, “progenitor” experiment). I veri-

fied eight deletion products by Sanger sequencing across the deletion junction (Fig. 5.1e). Only wild-type-size products were detected in the remaining clones and none of the 96 control clones exhibited any deletion bands (data not shown).

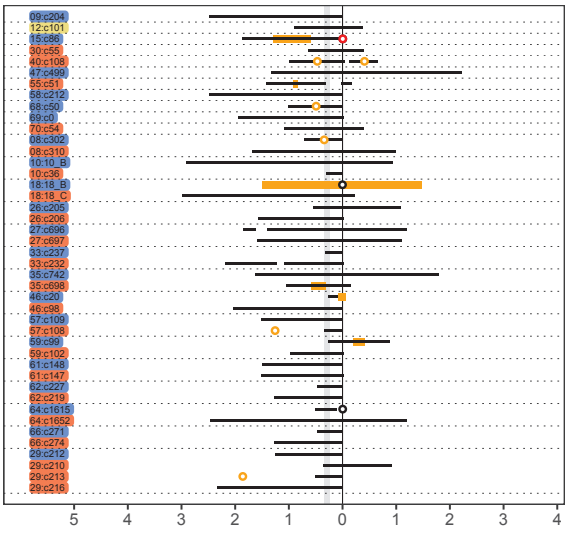
Observed frequency of deletions per clone was identical to the expected rate (36%), given 20% probability of each allele sustaining a deletion (as at the hemizygous *PigA* locus in mouse ES cells mutagenized with the exonic guide). However, the range covered by genotyping PCR was much smaller than in ES cells (5 kb vs 16 kb). While this result confirms large deletions are common, it also suggests that the real frequency is higher than expected, which may be a locus or cell-specific difference.

I have attempted to test this hypothesis by replicating the results at the *Cd45* locus (*Ptprc*) in an F1 cross combining two isoforms (*Cd45.1* and *Cd45.2*), whose expression can be distinguished through specific antibody staining. Initial experiments indicated that double knock-out is lethal in progenitor cells, as this population appeared only transiently in culture and did not form colonies when isolated by FACS (data not shown). Targeting *PigA* in progenitor cells failed to result in ablation of FLAER staining by day 11 post-delivery, at which point the experiment was terminated. It is possible that the effect would have been observed later. I decided not to target the *Cd9* locus, as without near 100% electroporation efficiency loss of Cd9 expression induced by intronic gRNAs would be very low. Furthermore, re-assessment of Cd9 expression status in outgrown colonies would not be possible using flow cytometry due to very low cell numbers. An immunofluorescence procedure could have been developed for that purpose.

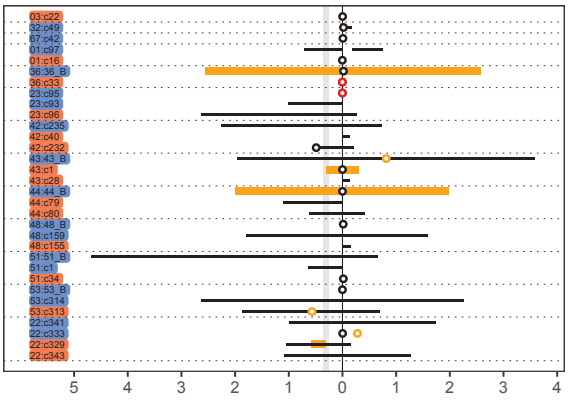
5.2.3 Insertions

Insertions (incl. duplications and inversions), defined as $\geq$10 bp fragments, which did not map in a linear fashion to the mutagenized locus, were present in 7-29% of resolved alleles from *PigA*, *PIGA* and *Cd9* loci. In almost all samples the most common origin of insertions was the edited locus (~62% of all insertions). This category ranged in size from small duplications <20 bp templated right next to the deletion breakpoint to perfect inversions of 3.9 kb (Fig. 5.4c). Fragments of *E.coli* genomic DNA and transfected plasmids up to 5 kb were found at all three examined loci, regardless of whether transfection was transient or involved mobilization of the PiggyBac transposon (Fig. 5.8). Distal insertions from introns and repetitive elements (predominantly LINE) were also present in a few samples. Notably, identity of four insertions of 13-29 bp could not be established, suggesting one non-templated or a few stitched, short, templated insertions.

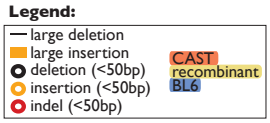
One of the alleles derived from PacBio sequencing of the edited *PigA* locus contained an insertion with a perfect match to four consecutive exons derived from the *Hmgn1* gene (Fig. 5.8a). It could represent a de novo insertion from the spliced and reverse-transcribed RNA, rather than from one of the pseudogenized forms of *Hmgn1*, as the pseudogenes diverge in sequence from the functional gene and thus from the observed insertion.

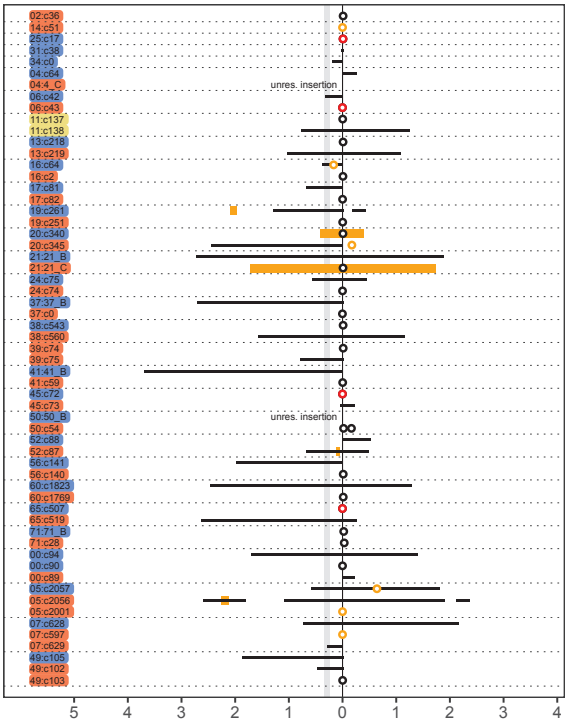


(a) Cd9^{low} clones.

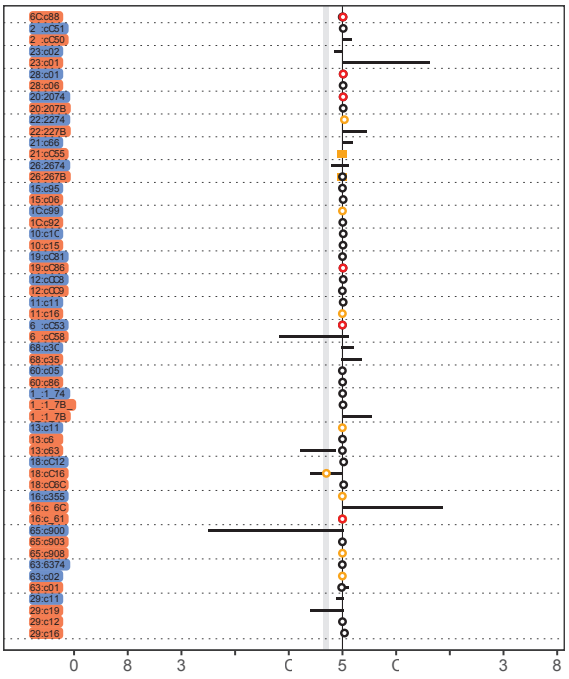


(c) Bimodal (mixed) clones.





(b) Wild-type clones, that were sorted for low expression ("low turned wild-type").



(d) Wild-type clones, that were sorted for wild-type expression ("true wild-type").

Figure 5.5: Alleles recovered by Sanger and PacBio sequencing from CAST/BL6 mouse ES cell clones mutagenized at the *Cd9* locus with the 3' intronic gRNA #1. PiggyBac constructs were stably delivered by transposition into Cas9-expressing cells. Gray shade represents exon 2. Dotted horizontal line separates clones. Clones are sorted by the number of alleles. Color behind allele names indicates strain of origin. Other display conventions as in Fig. 5.3.

Table 5.1: Results of PCR genotyping.

Experiment	gRNA	Gene	Primer pairs	Amplicon size [bp]	Target region	Sorted population	Clone expression	Total clones	≥1 del.	≥1 ins.	No amp.	% del.
cbbcs1	1	Cd9	5F/5R, 1F/1R	1063, 5554	intron	wt	wt	24	9	2	0	38%
cbbcs1	1	Cd9	5F/5R, 1F/1R	1063, 5554	intron	low	wt	14	6	3	0	43%
cbbcs1	1	Cd9	5F/5R, 1F/1R	1063, 5554	intron	low	low	10	5	0	1	50%
cbbcs1	35	Cd9	5F/5R, 1F/1R	1063, 5554	exon	wt	wt	24	7	1	0	29%
cbbcs1	35	Cd9	5F/5R, 1F/1R	1063, 5554	exon	medium	wt	20	7	0	0	35%
cbbcs1	35	Cd9	5F/5R, 1F/1R	1063, 5554	exon	medium	loss	4	2	0	0	50%
cbbcs1	35	Cd9	5F/5R, 1F/1R	1063, 5554	exon	loss	loss	24	4	3	1	17%
cbbcs1	80	Cd9	6F/6R, 3F/3R	1266, 5865	exon	medium	medium	24	7	0	0	29%
cbbcs1	80	Cd9	6F/6R, 3F/3R	1266, 5865	exon	loss	loss	24	10	1	0	42%
cbbcs1	86	Cd9	5F/5R, 1F/1R	1063, 5554	intron	wt	wt	24	8	1	0	33%
cbbcs1	86	Cd9	5F/5R, 1F/1R	1063, 5554	intron	low	wt	2	1	0	0	50%
cbbcs1	86	Cd9	5F/5R, 1F/1R	1063, 5554	intron	low	low	22	17	1	0	77%
cbbcs3	1	Cd9	4F/4R, 1F/1R, 2F/2R	1263, 5554, 11968	intron	wt	wt	24	11	2	0	46%
cbbcs3	1	Cd9	4F/4R, 1F/1R, 2F/2R	1263, 5554, 11968	intron	low	bimod.	12	8	3	0	67%
cbbcs3	1	Cd9	4F/4R, 1F/1R, 2F/2R	1263, 5554, 11968	intron	low	wt	31	22	5	0	71%
cbbcs3	1	Cd9	4F/4R, 1F/1R, 2F/2R	1263, 5554, 11968	intron	low	low	29	25	1	2	86%
cbbcs3	86	Cd9	5F/5R, 1F/1R	1063, 5554	intron	wt	wt	24	10	1	0	42%
cbbcs3	86	Cd9	5F/5R, 1F/1R	1063, 5554	intron	low	bimod.	11	9	3	0	82%
cbbcs3	86	Cd9	5F/5R, 1F/1R	1063, 5554	intron	low	wt	29	16	4	0	55%
cbbcs3	86	Cd9	5F/5R, 1F/1R	1063, 5554	intron	low	low	32	28	3	0	88%
progenitor	311	GFP	1F/2R, 1F/1R, 2F/2R	1314, 2994, 3507	exon	neg	N/A	96	35	0	2	36%
cherry/gfp	33	GFP	1F/1R, 1F/3R	972, 2291	exon	neg	neg	89	11	4	13	12%
cherry/gfp	34	mCherry	2F/2R, 2F/3R	1258, 2968	exon	neg	neg	46	3	3	12	7%
cherry/gfp	34	mCherry	2F/2R, 2F/3R	1258, 2968	exon	neg	pos	2	0	0	0	0%

Cells were edited with indicated guides, sorted for different gene expression levels ("Sorted population"), single cell cloned and reassessed for gene expression levels ("Clone expression"). **bimod.** - bimodal, **≥1 del.** - one or more deletion amplicons observed, **≥1 del.** - one or more insertion amplicons observed, **No amp.** - no amplicons, **% del.** fraction of clones with deletions amplicons.

5.2.4 Non-contiguous lesions

Notably, 13% of all alleles detected in single cell clones (56/428) contained additional lesions (SNPs, indels, large deletions and insertions) that were non-contiguous with the lesion at the cut site (Fig. 5.8b, c and d). This number is likely an underestimate due to stringent filtering of such variants at the *Cd9* locus (see Methods) and due to limited range of Sanger sequencing at the *PigA* and *PIGA* loci. For about 30% of non-contiguous lesions (17/56), the only exonic lesion detected was non-contiguous with the cut site. Furthermore, I observed alleles in which the intronic gRNA caused an inversion of a region containing the exon (Fig. 5.8c). Had the assessment been limited to the immediate vicinity of the cleavage site, such alleles would have been misclassified as wild type, and their phenotypic consequences would have been wrongly called.

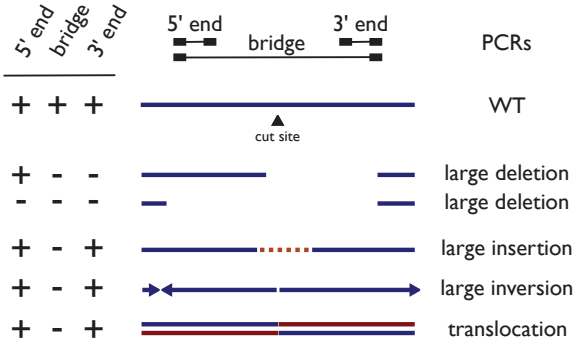


Figure 5.6: Results of "diagnostic" PCRs and their interpretations. Blue arrowheads indicate relative orientation of genomic fragments.

5.2.5 Unexpected genotypes of inconsistent clones

In mouse ES cells edited at the monoallelic *PigA* locus and sorted for loss of gene expression, I expected every clone to yield exactly one allele with a lesion overlapping the exon. One clone yielded two alleles, likely a result of a picking two closely growing colonies. Only seven out of remaining 164 clones did not contain a lesion overlapping the nearest exon (three were wild-type around the cut sites and four contained cut site, local lesions). They likely contained lesions in other exons or rearrangements outside of the amplified area that could ablate gene expression (e.g. large inversions containing the exon, insertions interfering with splicing, translocations within the gene). Since expression status of these clones was not ascertained after colony outgrowth, some of them could also be PigA proficient.

In ten cases, it was not possible to recover any product spanning the exon, even with a long-range PCR (16 kb). To understand this class of events, I performed additional, "diagnostic" PCRs targeting each end of the *PigA* locus (Fig. 5.1a, gray primers). In five cases, just one end or neither end of the locus could be amplified, suggesting a larger deletion. In the remaining five cases, both ends were amplified. Since no product connecting the two ends could be obtained, these are likely to be translocations, large inversions or large insertions (Fig. 5.3 and 5.6).

In female RPE1 cells edited at the *PIGA* locus and sorted for loss of gene expression, I expected every clone to yield exactly two alleles. At least one of them, presumably on the active chromosome X, should contain a lesion overlapping the nearest exon. No clone had more than two alleles and all clones with exactly two alleles had at least one exon-overlapping lesion, as expected. However, in about 32% of clones (14/44) only one allele was detected with PCR up to 12 kb. This could be due to a larger rearrangement (translocation, large deletion, insertion or inversion), which would explain loss of PIGA expression. Alternatively, five of the fourteen clones, in which an exon overlapping lesion was detected, could be monosomic or homologous for these lesions (there was no variants distinguishing the homologs). Therefore, the frequency of undetected alleles can range from 10% to 16% (9 or 14 alleles out of 88). The lower bound of this range is consistent with the rate of 8% (9/117) in mouse ES cell clones mutagenized with intronic gRNAs at the *PigA* locus, considering a slightly longer-range PCR was used (16 kb vs 12 kb). Higher rate could indicate a locus or cell-specific difference.

Clones derived from cells edited at the *Cd9* locus could be broadly classified into $Cd9^{low}$

and Cd9-positive (ie. bimodal, "low turned wild-type" and "true wild-type" clones; see chapter 4, Fig. 4.4c). The haplosufficient nature of the *Cd9* gene is demonstrated by the fact that I could detect at least one intact exon 2 in each one of the 67 Cd9-positive clones. Conversely, almost all $Cd9^{low}$ clones (25/26) had exon overlapping lesions in all detected alleles. The single exception contained an intronic insertion with a polyA signal. Furthermore, gene dosage could largely explain the difference between "true wild-type" and "low turned wild-type" clones. The first group usually contained at least two functional exon 2s (22/24), while the second group usually had exactly one (27/30), consistent with their 50% lower Cd9 expression (Fig. 5.5).

For experiments at the *Cd9* locus, I used mouse ES cells derived from an F1 cross between BL6 and CAST mouse strains, which allowed me to distinguish the homologous chromosomes. In no case was the repair outcome identical between homologs within a clone, despite 15 alleles reoccurring between clones. Just over half of the mutagenized clones (52/93) contained precisely one BL6 and one CAST allele, as expected. Notably, in 18 clones only one allele was detected with PCR spanning up to 12 kb, potentially due to a larger rearrangement (translocation, large deletion, insertions or inversions), monosomy or LOH. Some of the wild-type BL6 alleles removed as feeder-derived could be the missing alleles. An abnormal number of alleles (two from the same strain or more than two in total) was found in 21 clones, which could have resulted from picking two closely growing colonies, large duplication, repair events happening during clone outgrowth or aneuploidy (spontaneous or induced by Cas9 cutting).

Two clones contained recombinant BL6-CAST alleles (Fig. 5.9). In one case, a LOH event distal to the breakpoints converted part of the CAST allele to BL6. In another case, the BL6-CAST crossover boundary did not coincide with the breakpoint. I concluded that the creation of these alleles likely involved interhomolog strand invasion as they cannot be explained by a simple rejoining of the resected ends of two broken chromosomes.

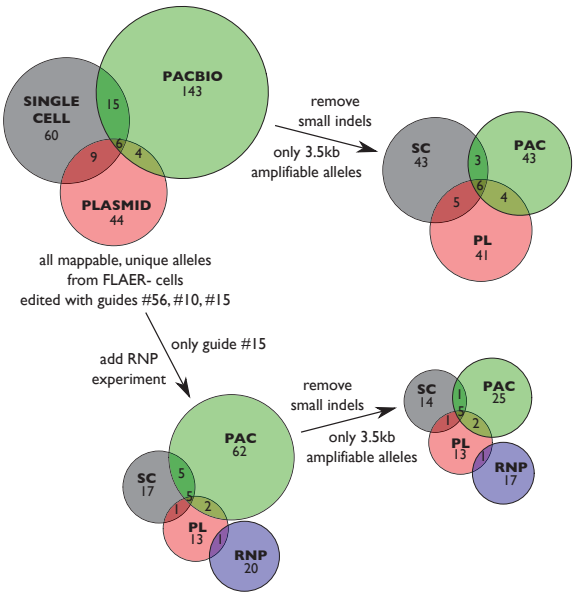


Figure 5.7: Overlap between unique *PigA* alleles derived using different methods. “PacBio” and “Single cell” refer to alleles shown in Fig. 5.2, 5.3a, 5.3c and 5.3d. “Plasmid” alleles were derived in the same experiment from subcloned PCR amplicons. “RNP” alleles were derived in an independent experiment only using guide #15 (Fig. 5.3b).

5.2.6 Diversity of resolved alleles at the *PigA* locus

To gauge the diversity of mutagenesis outcomes, I have compared unique, sequence resolved alleles that were derived in one experiment from PigA-deficient mouse ES cells edited with intronic gRNAs #10, #15 and an exonic gRNA #56 using following three methods:

- PacBio sequencing of 5.5 kb PCR products amplified from bulk DNA (2F/2R primer pair in Fig. 5.1a; results in Fig. 5.2)
- Sanger sequencing of single cell clones, with PCR product up to 16 kb in size (all primer pairs in Fig. 5.1a; results in Fig. 5.3a, 5.3c, 5.3d)
- Sanger sequencing of individual 3.5 kb PCR products amplified from bulk DNA

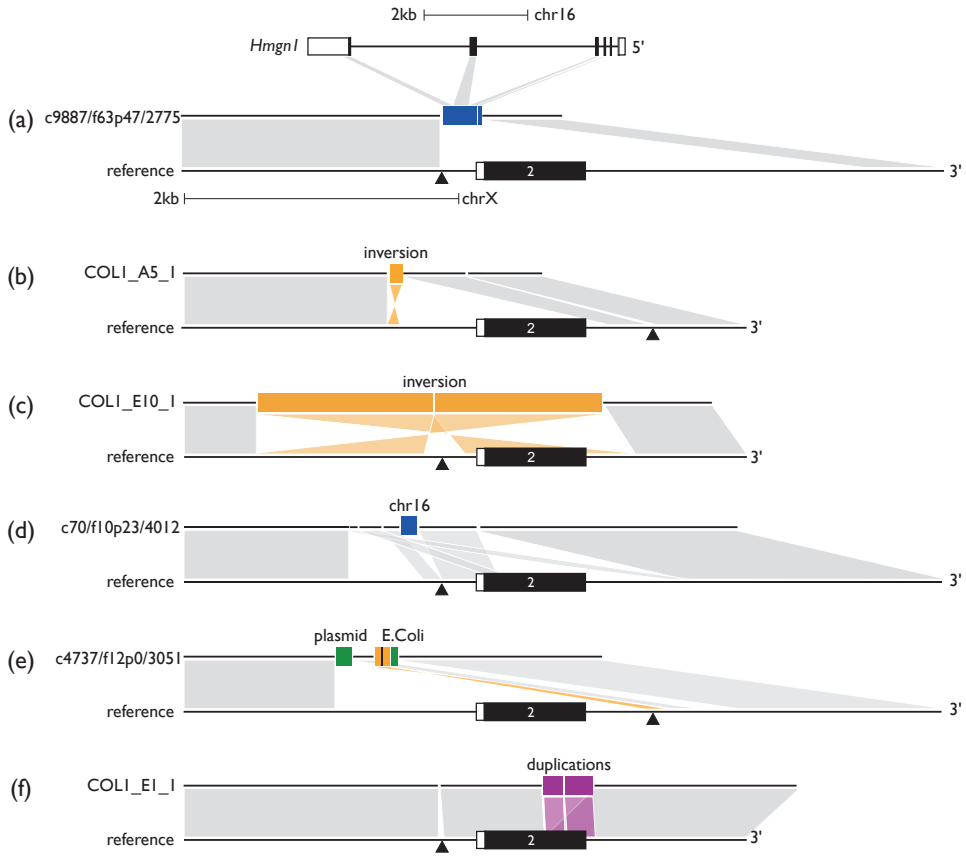


Figure 5.8: Examples of alleles. The bottom diagram of each panel represents the *PigA* reference allele around exon 2, the diagram immediately above shows the structure of the sequenced allele. Black horizontal line: direct reference match; orange bar: inversion; blue bar: insertion from another part of the genome; violet bar – duplication; black arrowhead: gRNA target site. Gray, orange and violet shadows represent, respectively, direct, inverted and duplicated match between the reference and the sequenced allele. Lack of shadow at the reference locus represents a deletion in the sequenced allele. (a) Putative insertion from a reverse transcribed RNA. The top diagram line shows the genomic structure of *Hmgn1*; note the scale differs from that of *PigA* gene. (b) Exonic lesion non-contiguous with the cut site. (c) Inversion of a region containing the exon. (d) "Scrambled" allele with insertion from chromosome 16. (e) Combined deletion, local inversion and insertion from *E.Coli* genome. (f) Duplication of a region containing the exon.

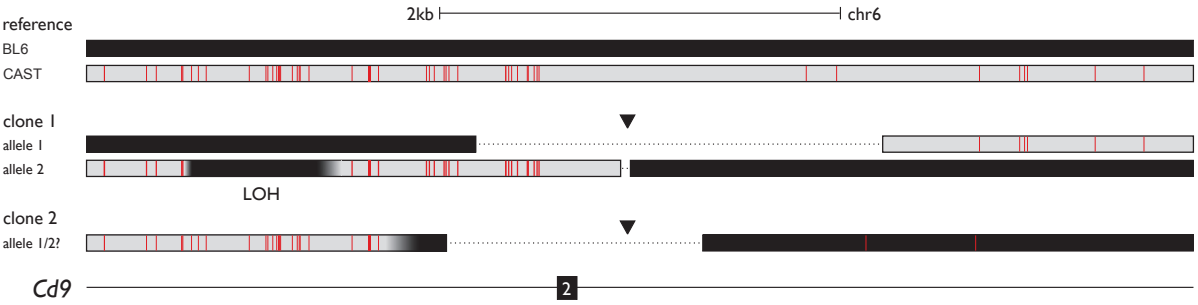


Figure 5.9: Recombinant *Cd9* alleles. Two of the sequenced single cell clones contained alleles indicative of a cross-over event between the homologous chromosomes. Red vertical bars in CAST allele (gray bar) indicate positions of sequence divergence from the BL6 reference genome (black bar), dotted black line indicates missing sequence (deletion), thin black line indicates an intron. LOH: loss of heterozygosity.

and cloned into plasmid vectors (1F/1R primer pair in Fig. 5.1a)

Clustering of PacBio reads from PigA-deficient samples yielded 168 unique alleles. The majority of the alleles recovered from single cell clones and plasmid cloned products were unique (90/130 and 63/75, respectively; wild-type alleles excluded). In total 281 unique alleles were recovered by the three methods, only 31 of which (11%) were shared between two or three methods (Fig. 5.7).

To make the comparison more reliable, I removed alleles which could not be recovered with the 3.5 kb primer pair (1F/1R) and small indels (<10 bp), which were depleted from PacBio clusters due to method-specific biases described in subsection 5.2.1. Out of the remaining 145 unique alleles only 18 (12%) were detected with more than one method (Fig. 5.7).

I also compared the alleles in this experiment with ones derived by single cell cloning of RNP mutagenized cells, keeping only the alleles mutagenized by the intronic gRNA #15. The only overlap observed was between one out of the 21 unique RNP alleles and one allele in the plasmid cloned group. I concluded that the large allelic diversity of mutagenic outcomes may be difficult to describe exhaustively using sequencing based methods.

5.2.7 Diversity of deletion fingerprints at the *PigA* locus

Diversity of deletion outcomes can be visualized by resolving PCR products from pools of edited cells on an agarose gel. If enough cells were used in each experiment to avoid stochastic undersampling of different deletion outcomes, the ladder-like pattern corresponding to different deletion sizes ("deletion fingerprint") should be similar between biologically independent replicates.

I repeated the original experiment four times using intronic gRNA #15 in two mouse ES cell lines, the original JM8 (also transfected with a PiggyBac Cas9 vector) and its subclone expressing Cas9 from a single-copy lentiviral transgene. I sorted PigA-deficient cells and performed 3.5 kb PCR (1F/1R primer pair) on bulk extracted DNA. I assumed 40% transfection and stable transposition efficiency of the gRNA plasmid, 15% frequency of PigA loss due to intronic mutagenesis (Fig. 4.2b), 20% plating efficiency of mouse ES cells and ability to amplify 80% of PigA-deficient alleles using the 3.5 kb PCR (35/43 among single cell clones). Starting with 1.5 million cells this translates into a transfection bottleneck of 15,000 individual cells with detectable deletions in the Cas9 expressing line (Fig. 5.10a). After antibiotic selection and population outgrowth, I sorted one million cells from each sample, ensuring more than 60x coverage of the bottleneck.

I asked whether sampling of 15,000 unique cells bearing >200 bp deletions is enough to cover the diversity of the possible deletion outcomes. I assumed that the sorting step (with 60x coverage) did not reduce this initial diversity. However, the PCR reaction itself could introduce stochastic noise into the procedure, if too few products were sampled from the pool of extracted genomic DNA in each individual reaction. In order to ensure that PCR was not the limiting factor I have performed a series of technical duplicate PCR reaction starting with 250, 2,500 and 12,500 DNA copies (80% of which should be possible to amplify with the “short” PCR) and compared their "deletion fingerprints" (Fig. 5.10). Sampling 250 copies led to loss of technical reproducibility, as PCR duplicates differed substantially. With 2,500 copies, the diversity of the biological replicate #4 was preserved, revealing it to be the least complex in the set. Sampling 12,500 copies preserved diversity of all replicates. Although some similarities could be observed across biological replicates and between the two cell lines in the same biological replicate, my general conclusion is that sampling ~15,000 cells did not sufficiently cover the diversity of deletion alleles.

5.3 Discussion

Using long-range PCR, I have genotyped in excess of 850 single cell clones mutagenized with Cas9,

Table 5.2: Summary classification of alleles.

Gene	gRNA	Target	Expr.	Method	Indel	Deletion >50 bp	Insertion >10 bp	Multi -Lesion	Intact Exon	WT	Total alleles	Total clones
PigA	15	intron	neg	RNP	0	22	4	3	1	0	24	24
PigA	15	intron	neg	PiggyBac	0	40	9	10	1	1	48	48
PigA	10	intron	neg	PiggyBac	0	39	7	12	3	0	45	45
PigA	56	exon	neg	PiggyBac	32	10	6	1	2	2	48	47
PIGA	274	intron	neg	transient	7	19	4	2	12	3	32	16
PIGA	275	intron	neg	transient	12	16	6	5	15	0	34	19
PIGA	276	intron	neg	transient	6	5	4	0	8	2	22	9
Cd9	1	intron	low	PiggyBac	0	42	9	10	1	0	43	26
Cd9	1	intron	bimod.	PiggyBac	8	20	6	4	14	0	32	13
Cd9	1	intron	l-wt	PiggyBac	24	30	9	5	33	0	59	30
Cd9	1	intron	wt	PiggyBac	33	18	3	4	50	0	55	24

Expr.: expression class; **bimod.:** bimodal; **l-wt:** low turned wild-type; **indel:** small deletion and/or insertion only (<50 bp); **Intact exon:** intact exon in the correct orientation. Categories are not mutually exclusive.

about 300 of which were also sequenced at the mutagenized locus using Sanger and PacBio technologies. The results revealed a pervasive presence of large deletions (50 bp - 9.5 kb), which explains frequent loss of gene expression upon intronic cutting. Many complex rearrangements of the locus, including large insertions, inversion, translocation between homologs and non-contiguous lesions were also discovered.

5.3.1 Consequences of large deletions

Large deletions could be pathogenic in gene therapeutic context. Given that a target locus would presumably be transcriptionally active, such mutations could juxtapose it to the nearest oncogene, initiating neoplasia. A deletion inactivating a nearby tumor suppressor gene could predispose the cell to become cancerous, even if only one copy is affected (Santarosa and Ashworth, 2004). The effect might not be immediately obvious, as the lesions may constitute a carcinogenic first "hit". This is especially true for stem cells and progenitors, which have a long replicative lifespan and may become neoplastic with time. This would be similar to the activation of *LMO2* by pro-viral insertion in some of the early gene-therapy trials, which caused cancer in these patient (Hacein-Bey-Abina et al., 2003).

The closer the target site is to a cancer-driver gene, the higher the risk posed by deletions and other local rearrangements. I have not gathered enough unbiased data at any locus to accurately describe the frequency of "complex" lesions as a function of distance from the cut site. However, the gene expression data at the *PigA* locus comes close. A simple exponential model can be fitted using exon 2 proximal (100-500 bp) and distal (~2 kb) gRNAs. I assumed that loss of PigA expression caused by gRNAs close to exon 2 is exclusively due to damage to exon 2 (and not exon 1 or exon 3) and that the two gRNAs in the middle of intron 2 confer double the risk by affecting both exon 2 and 3 (data not shown). As a crude reality check, I asked if the model correctly predicts the tail of the distribution - the largest deletion in the Sanger sequencing dataset (which is 9.5 kb in total, 6.6 kb in one direction). Given 117 intronically edited alleles from PigA-deficient cells were tested, the model indicates on average 1.43 such lesions (or larger) should be found, which is consistent with reality.

The lesion frequency under this model halves with every kilobase of distance from the cut site. This implies that for every 100 million mutagenized cells (the scale of current gene therapeutic efforts), one lesion spanning 22.5 kb or more in one direction from the cut site would be expected,

on the average. While such calculations are subject to a very high statistical uncertainty and may not generalize to other loci, they could inform the design of future experiments with respect to investigated distances and numbers of cells.

5.3.2 Consequences of other complex lesions

Sequencing of single cell clones yielded large insertions, inversions, non-contiguous lesions, cross-overs and LOH events. Some of these were directly implicated in causing gene expression loss, notably inversions containing the exon, non-contiguous lesions within the exons and an intronic insertion containing polyA signal. Furthermore, the consequences of some of these lesions would have been underestimated, if only genotyping around the cut site was performed. This suggests that genotyping should not be limited to the immediate vicinity of the cut site and stresses the importance of careful phenotypic assessment, whenever possible.

The full extent of non-contiguous lesions is not known. Sanger sequencing was primarily performed to detect deletion breakpoints, resolve insertions and ensure integrity of the exon closest to the cut site, so more distal lesions could have been missed. I have observed some small, distal indels in alleles derived by PacBio sequencing of the *Cd9* locus. However, I decided to filter them out, as some of them consistently clustered at the ends of the read (indicating quality issues) and in low complexity regions (where accurate mapping turned out to be an issue). Such lesions could be investigated in the future using more reliable Sanger sequencing. As with large deletions, a quantitative description of the frequency of non-contiguous lesions as a function of distance from the cut site would be useful in gene therapeutic context.

Cas9-induced cross-overs would have minimal impact if products co-segregate on cell division (i.e. undergo a "z-segregation"). If instead they segregate away from each other ("x-segregation"), a cross-over would result in chromosome-scale LOH, which could uncover recessive alleles. If tumor suppressor genes are affected, this could initiate cancer.

Analysis of single cell clones has indicated presence of aneuploidies and alleles that could not be fully resolved, such as translocations. These events are often observed in cancers, due to their ability to juxtapose active promoters and oncogenes, amplify oncogenes and reduce the copy number of tumor suppressors. I have not performed a more detailed analysis of these clones, but copy-number screening by qPCR or digital droplet PCR, karyotyping to detect translocations and SNP array genotyping for large scale deletions and LOH events are warranted. Furthermore, it would be crucial to establish the causal relationship between Cas9 mutagenesis and aneuploidies, as ES cells in culture are known to acquire aneuploidies spontaneously.

Failure to detect one of the two lesions at an autosomal locus in a single cell clone (or a founder animal) can be easily mistaken for a homozygous lesion. While in the context of animal mutagenesis such mistake should be detected when animals fail to breed true (as discussed in Shin et al., 2017), it can significantly influence the interpretation of experiments using single cell clones, whose alleles cannot be easily isolated.

These considerations formed the basis of debate, in which I was involved, on the interpretation of a particular human embryo editing study. In that study, researchers used Cas9 to induce a DSB on the paternal allele in human zygotes and observed only the maternal allele in some blastomers isolated from the multicellular embryo three days later. In absence of further evidence, it was concluded that the maternal allele served as a template for the repair of the paternal allele, a process termed "interhomolog repair" (Ma et al., 2017). This conclusion has been challenged by two groups as equally consistent with a failure to detect the paternal allele due to destruction of primer binding sites (Adikusuma et al., 2018; Egli et al., 2018). One of these groups included data which showed edited mouse embryos exhibit high levels of large deletions. Reply to this criticism reported no deletions with PCR spanning up to

10 kb and established that at least some blastomers carry the expected heterozygous patterns of SNPs flanking the target site, which supports the original conclusion (Ma et al., 2018). Another group independently studying interhomolog repair in mouse embryos also carried out the prescribed checks (long-range PCRs, copy-number qPCR) and failed to observe large losses of genetic material (Wilde et al., 2018). With some other groups reporting detection of large deletions in human embryos and in differentiated animal tissues edited in vivo (personal communication), it remains to be established which conditions enable creation of complex lesions (also see the Discussion chapter).

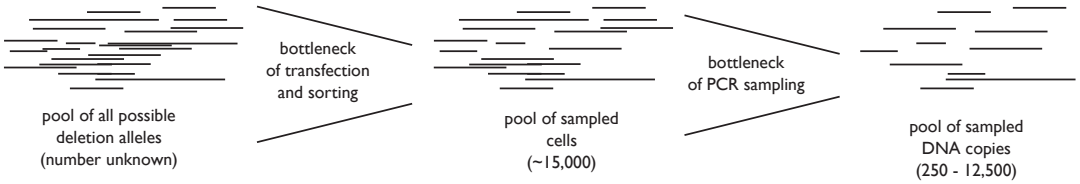
5.3.3 Stochasticity of large deletions

Indels induced by any gRNA are highly non-random, with a few indels of particular sizes forming a stable "indel profile". Such profiles are hypothesized to be related to local microhomologies guiding the repair process. I speculated an analogous "deletion profile" exists, potentially also guided by homologies or larger scale secondary structure of the DNA. Ladder pattern observed by resolving amplicons from long-range PCRs on pools of mutagenized cells initially seemed to confirm this hypothesis. However, the observed "profiles" differed between biological replicates.

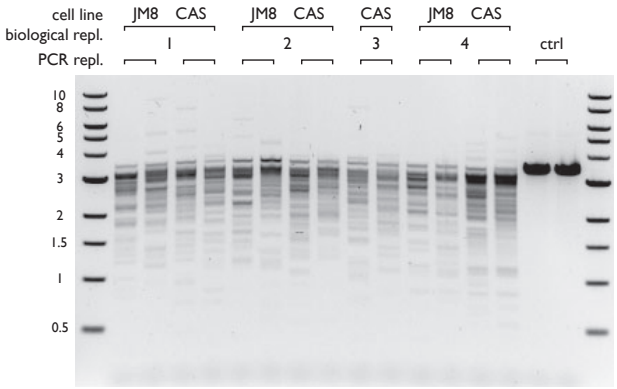
Two possible explanations exists for the lack of reproducibility of "deletion profiles". One is that the potential diversity of induced deletions far outstrips the number of transfected cells with deletion outcomes. This would lead to stochastic undersampling of deletion events in each transfection replicate, resulting in a "noisy" profile. If this model is correct, sampling more cells should eventually reduce the noise between biological replicates. This could be achieved at a scale by employing a non-leaky, inducible gRNA and Cas9 system. Another explanation could be clonal expansion due to stochastic genomic instability. In this model, cells which acquired a mutation that makes them grow faster (e.g. chromosome 8 triploidy) selectively amplify the Cas9-induced deletion they harbor. If this model is correct, a more karyotypically stable cell line should behave more predictably. If such genomic instability is independent of Cas9 mutagenesis, then even wild-type cells will exhibit strong clonal effects, which could be tested by random barcoding. Regardless, my results revealed a source of noise that needs to be taken into account when investigating "deletion profiles".

5.3.4 Other considerations

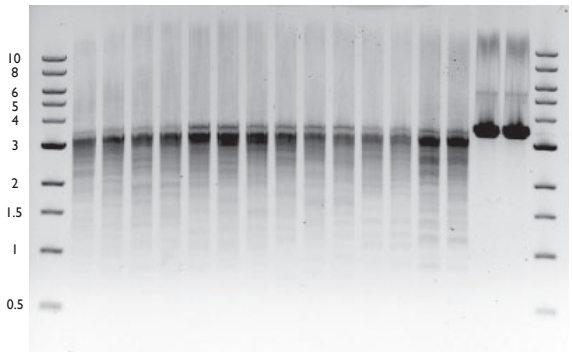
Most of the "low turned wild-type" clones edited at the *Cd9* locus had exactly one exon-overlapping and one non-overlapping lesion, as opposed to "true wild-type" clones, most of which did not have any exon-overlapping lesions. Although the difference between these populations was not immediately apparent in bulk cultures (Fig. 4.4a), improved culturing protocols and use of single cell clone controls could potentially allow systematic quantification and isolation of (or at least enrichment for) cells with monoallelic "complex" lesions at the *Cd9* locus.



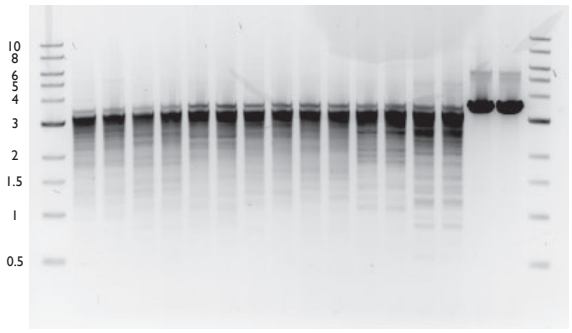
(a) Experimental considerations.



(b) Deletion fingerprint - 250 copies per reaction.



(c) Deletion fingerprint - 2,500 copies per reaction.



(d) Deletion fingerprint - 12,500 copies per reaction.

Figure 5.10: *PigA* locus was mutagenized using 5' intronic guide (#15) in biological quadruplicate. Duplicate PCR reactions spanning the cut site performed on DNA extracted from the bulk of PigA-deficient cells and resolved on an agarose gel (product size 3,500 bp, 1F/1R primer pair, Fig. 5.1a). JM8 – original mouse ES cell line (transfected with gRNA and Cas9 PiggyBac vectors); CAS – JM8 subclone stably expressing Cas9 (transfected only with a gRNA PiggyBac vector). Ladder scale is in kilobases.

Table 5.3: Genomic primer pairs.

Gene	Name	Sequence	Chr	Start	End	Strand
PigA	1F	CTTATGGGATGTACTGGGTCACTAG	X	164421324	164421349	+
PigA	1R	CACCCCAGAAAATGTAACTGAGTTC	X	164424799	164424824	-
PigA	2F	CTTTCATTTGGTTCATTATTTCTGTTCTTATC	X	164420461	164420493	+
PigA	2R	CCTTAACTCAAGAGCTGAACTT	X	164425873	164425895	-
PigA	3F	TTCGACCAGTTTGCTCTAACTCTTA	X	164417878	164417903	+
PigA	3R	ATCAAAGTGTCCTCGAGTTAAT	X	164430740	164430762	-
PigA	4F	AAGCTCTTAAGAGGAAAGGCTACAA	X	164417360	164417385	+
PigA	4R	ATCACACCACAGCATTAGGA	X	164418508	164418528	-
PigA	5F	TAACAGGTCACATATAGGATTTGGG	X	164414904	164414929	+
PigA	6F	ATGTGGAAATCCTGTACCAGAAAGA	X	164429755	164429780	+
PigA	6R	AACTGATTATCTGACCTTCCCT	X	164423503	164423525	-
PigA	7F	AGGAGACTGAGGCCAGGAATAT	X	164421983	164422005	+
PIGA	1F	CGGTTCACATGTTCCTGATTAAAGAA	X	15328961	15328987	+
PIGA	1R	GTGGTCGAGAATTTTACGGTAATGT	X	15334958	15334983	-
PIGA	2F	CTTTCCCGAACTTCTTCCAAAATGA	X	15325931	15325956	+
PIGA	2R	AGGCAGGACACCATAATTAGAATCA	X	15337669	15337694	-
Cd9	1F	CTTTAGTGTCCTTTTGCACACTTCT	6	125474857	125474882	-
Cd9	1R	GGTATAACCAGTCCTTCTAGCACAT	6	125469328	125469353	+
Cd9	2F	CTGTCGTGAAATATTAGGAAAGGGC	6	125477789	125477814	-
Cd9	2R	AGTACCTCCCGTCTTGCTACC	6	125465846	125465867	+
Cd9	3F	ATCTGAAGAAGTCTCTCTGACCCTA	6	125467206	125467231	-
Cd9	3R	TCTTCTTTGGTGATTTGCTGATTCC	6	125461366	125461391	+
Cd9	4F	AGTTTTCTGGTGATTTTACCGCAAT	6	125472672	125472697	-
Cd9	4R	CCTTGTCAGAATGCTTTCTTGTCTT	6	125471434	125471459	+
Cd9	5F	ATCATTTGGCATCCTATTCAACACC	6	125473010	125473035	-
Cd9	5R	CTCCATCTCCATCCCCATTAATCTC	6	125471972	125471997	+
Cd9	6F	AGGTCTCAGTAAGTTAGCTCAAGTG	6	125464803	125464828	-
Cd9	6R	ATAAGGAGGTGTGATCAGTGGAAAA	6	125463562	125463587	+
Cas9-GFP	1F	AGAAACTGAAGAGTGTGAAAGAGC	-	-	-	+
Cas9-GFP	1R	CGTGCAATCCATCTTGTTCAATG	-	-	-	-
Cas9-GFP	2F	GGCGGCAGGAAGATTTTTACCC	-	-	-	+
Cas9-GFP	2R	GGGTGTTCTGCTGGTAGTGGT	-	-	-	-
cherry/gfp	1F	GTAAACGGCCACAAGTTCAGC	-	-	-	+
cherry/gfp	1R	GCTCAAGATGCCCCTGTTCT	-	-	-	-
cherry/gfp	2F	GGAGGATAACATGGCCATCATCAAG	-	-	-	+
cherry/gfp	2R	CTGATGCTCTTCGTCCAGATCA	-	-	-	-
cherry/gfp	3R	TTGACCTATTCTGGCATTGTAGACA	-	-	-	-

Genomic position is given with respect to the GRCm38 or GRCh38.